



22883

PATENT TRADEMARK OFFICE

103.1072.01

1 This application is submitted in the name of the following inventors:

2	3 <i>Inventor</i>	<i>Citizenship</i>	<i>Residence City and State</i>
4	Blake LEWIS	USA	Palo Alto, California
5	John EDWARDS	USA	Sunnyvale, California
6	Srinivasan VISWANATHAN	India	Fremont, California

7

8 The assignee is *Network Appliance, Inc.*, a California corporation having an

9 office at 495 East Java Drive, Sunnyvale, CA 94089.

10

11 TITLE OF THE INVENTION

12

13 Instant Snapshot

14

15 CROSS-REFERENCE TO RELATED APPLICATION

16

17 This application is a continuation-in-part of U.S. Patent Application Serial

18 No. 09/642,061, Express Mail Mailing No. EL 524 780 239 US, filed August 18, 2000,

19 in the name of the same inventors, attorney docket number 103.1035.01, titled "Instant

20 Snapshot".

BACKGROUND OF THE INVENTION

1. *Field of Invention*

This invention relates to data storage systems.

2. *Related Art*

Snapshots of a file system capture the contents of the files and directories in a file system at a particular point in time. Such snapshots have several uses. They allow the users of the file system to recover earlier versions of a file following an unintended deletion or modification. The contents of the snapshot can be copied to another storage device or medium to provide a backup copy of the file system; a snapshot can also be copied to another file server and used as a replica. The WAFL (Write Anywhere File Layout) file system includes a copy-on-write snapshot mechanism. Snapshot block ownership in WAFL has been recorded by updating the block's entry in a blockmap file, which is a bitmap indicating which blocks are in-use and which are free for use.

One problem with the prior art of creating snapshots is that the requirement for additional file system metadata in the active file system to keep track of which blocks snapshots occupy. These methods are inefficient both in their use of storage space and in the time needed to create the snapshots.

1
2 A second problem with earlier snapshot implementations, was the time
3 consuming steps of writing out a description of the snapshot state on creation and re-
4 moving it on deletion.

5
6 A third problem with earlier copy-on-write mechanisms, was the required
7 steps consumed a considerable amount of time and file system space. For example, some
8 systems, such as those supplied with DCE/DFS, include a copy-on-write mechanism for
9 creating snapshots (called "clones"). The copy-on-write mechanism was used to record
10 which blocks each clone occupied. Such systems require a new copy of the inode file
11 and the indirect blocks for all files and directories are created when updating all of the
12 original inodes.

13
14 Accordingly, it would be advantageous to provide an improved technique
15 for more quickly and efficiently capturing the contents of the files and directories in the
16 file system at a particular point in time. This is achieved in an embodiment of the inven-
17 tion that is not subject to the drawbacks of the related art.

18 19 SUMMARY OF THE INVENTION

20
21 The invention provides an improved method and apparatus for creating a
22 snapshot of a file system.

1
2 In a first aspect of the invention, the file system uses the fact that each
3 snapshot includes a representation of the complete active file system as it was at the time
4 the snapshot was made, including the blockmap of disk blocks indicating which ones are
5 free and which ones are in use (herein called the "active map"). Because a record of
6 which blocks are being used by the snapshot is included in the snapshot itself, the file
7 system can create and delete snapshots very quickly. The file system uses those recorded
8 blockmaps (herein called "snapmaps") as a source of information to determine which
9 blocks cannot be reused because those blocks are being used by one or more snapshots.

10
11 In a second aspect of the invention, the file system uses that fact that it need
12 only maintain a more limited blockmap of those disk blocks in use by the active file sys-
13 tem, and a summary map of those disk blocks in use by one or more snapshots. The
14 summary map can be computed from the snapmaps as the logical inclusive-OR of all the
15 snapmaps. Because the file system need not maintain multiple bits of in-use/free data for
16 each block, it uses the active map in conjunction with the summary map to determine
17 whether blocks are in-use or free.

18
19 In a third aspect of the invention, the file system makes use of the fact that
20 the summary map need not be updated every time a block is allocated or freed. Accord-
21 ingly, the file system updates the summary map only (1) when a snapshot is deleted, and
22 then only in a background operation, (2) on demand for areas for which write allocation

1 is about to be performed, and (3) periodically in a background operation for selected por-
2 tions of the summary map. These background operations are preferably performed con-
3 currently with other file system operations.

4
5 Information is stored in a persistent storage medium accessible by the file
6 system, to provide for resumption of operation following a reboot operation. For exam-
7 ple, in a preferred embodiment, relevant information is stored in the file system "fsinfo
8 block" for each snapshot, to indicate whether the summary file needs to be updated using
9 that snapshot's snapmap information as a consequence of its creation or deletion. When a
10 block is freed in the active file system, the corresponding block of the summary file is
11 updated with the snapmap from the most recently created snapshot, if this has not already
12 been done. An in-core bit map records the completed updates to avoid repeating them
13 unnecessarily. This ensures that the combination of the active bitmap and the summary
14 file will consistently identify all blocks that are currently in use. Additionally, the sum-
15 mary file is updated to reflect the effect of any recent snapshot deletions when freeing a
16 block in the active file system. This allows reuse of blocks that are now entirely free.
17 After updating the summary file following a snapshot creation or deletion, the corre-
18 sponding bit in the fsinfo block is adjusted.

19
20 In a fourth aspect of the invention, the algorithm for deleting a snapshot in-
21 volves examining the snapmaps of the deleted snapshot and the snapmaps of the next
22 oldest and next youngest snapshot. A block that was used by the deleted snapshot but is

1 not used by its neighbors can be marked free in the summary file, as no remaining snap-
2 shot is using it. However, these freed blocks cannot be reused immediately, as the snap-
3 map of the deleted snapshot must be preserved until summary updating is complete.
4 During a snapdelete free blocks are found by using the logical OR of the active bitmap,
5 the summary file, and the snapmaps of all snapshots for which post-deletion updating is
6 in progress. In other words, the snapmap of the deleted snapshot protects the snapshot
7 from reuse until it is no longer needed for updating.

8
9 In the preferred embodiment, the invention is operative on WAFL file sys-
10 tem. However, it is still possible for the invention to be applied to any computer data
11 storage system such as a database system or a store and forward system such as cache or
12 RAM if the data is kept for a limited period of time.

13 BRIEF DESCRIPTION OF THE DRAWINGS

14
15
16 Figure 1 shows a block diagram of a system for an instant snapshot.

17
18 Figure 2 shows a block diagram of an instant snapshot.

19
20 Figure 3 shows a flow diagram of a method for creating a snapshot.

21
22 Figure 4 shows a flow diagram of a method for updating a summary map.

Figure 5 shows a block diagram of copy-on-write maintenance of the active map.

INCORPORATED DISCLOSURES

The inventions described herein can be used in conjunction with inventions described in the following applications:

- U.S. Patent Application Serial No. 09/642,063, Express Mail Mailing No. EL524781089US, filed August 18, 2000, in the name of Blake LEWIS, attorney docket number 103.1033.01, titled "Reserving File System Blocks".
- U.S. Patent Application Serial No. 09/642,062, Express Mail Mailing No. EL524780242US, filed August 18, 2000, in the name of Rajesh SUNDARAM, attorney docket number 103.1034.01, titled "Dynamic Data Storage".
- U.S. Patent Application Serial No. 09/642,066, Express Mail Mailing No. EL524780256US, filed August 18, 2000, in the name of Ray CHEN, attorney docket number 103.1047.01, titled "manipulation of Zombie Files and Evil-Twin Files".

- 1 • U.S. Patent Application Serial Number 09/642,064, in the names of Scott
2 SCHOENTHAL, Express Mailing Number EL524781075US, titled "Persistent
3 and Reliable Delivery of Event Messages", assigned to the same assignee, attorney
4 docket number 103.1048.01, and all pending cases claiming the priority thereof.

5
6 and

- 7
8 • U.S. Patent Application Serial Number 09/642,065, in the names of Douglas P.
9 DOUCETTE, Express Mailing Number EL524781092US, titled "Improved Space
10 Allocation in a Write Anywhere File System", assigned to the same assignee, at-
11 torney docket number 103.1045.01, and all pending cases claiming the priority
12 thereof.

13
14 DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

15
16 In the following description, a preferred embodiment of the invention is de-
17 scribed with regard to preferred process steps and data structures. However, those skilled
18 in the art would recognize, after perusal of this application, that embodiments of the in-
19 vention might be implemented using a variety of other techniques without undue experi-
20 mentation or further invention, and that such other techniques would be within the scope
21 and spirit of the invention.

1 *Lexicography*

2
3 As used herein, use of the following terms refer or relate to aspects of the
4 invention as described below. The general meaning of these terms is intended to be il-
5 lustrory and in no way limiting.

- 6
- 7 • fsinfo (File System Information Block) — In general, the phrase "file system in-
8 formation block" refers to one or more copies of a block known as the "fsinfo
9 block". These blocks are located at fixed locations on the disks. The fsinfo block
10 includes data about the volume including the size of the volume, volume level op-
11 tions, language and more.
 - 12
 - 13 • WAFL (Write Anywhere File Layout) — In general, the term "WAFL" refers to a
14 high level structure for a file system. Pointers are used for locating data. All the
15 data is included in files. These files can be written anywhere on the disk in chunks
16 of file blocks placed in data storage blocks.
 - 17
 - 18 • Consistency Point (CP) — In general, the term "CP" refers to a time that a file
19 system reaches a consistent state. When this state is reached, all the files have
20 been written to all the blocks and are safely on disk and the one or more copies of
21 redundant fsinfo blocks get written out. If the system crashes before the fsinfo

1 blocks go out, all other changes are lost and the system reverts back to the last CP.

2 The file system advances atomically from one CP to the next.

3
4 • Consistent State — In general, the phrase "consistent state" refers to the system
5 configuration of files in blocks after the CP is reached.

6
7 • Active file system — In general, the phrase "active file system" refers to the cur-
8 rent file system arrived at with the most recent CP. In the preferred embodiment,
9 the active file system includes the active map, the summary map and points to all
10 snapshots and other data storage blocks through a hierarchy of inodes, indirect
11 data storage blocks and more.

12
13 • Active map — In general, the phrase "active map" refers to a to a file including a
14 bitmap associated with the in-use or free status of blocks of the active file system.

15
16 • Snapshot — In general, the term "snapshot" refers to a copy of the file system.
17 The snapshot diverges from the active file system over time as the active file sys-
18 tem is modified. A snapshot can be used to return the file system to a particular
19 CP (consistency point).

20
21 • Snapmap — In general, the term "snapmap" refers to a file including a bitmap as-
22 sociated with the vacancy of blocks of a snapshot. The active map diverges from a

1 snapmap over time as the blocks used by the active file system change during con-
2 sistency points.

- 3
- 4 • Summary map — In general, the term "summary map" refers to a file including an
5 IOR (inclusive OR) bitmap of all the snapmaps.
6
 - 7 • Space map — In general, the term "space map" refers to a file including an array
8 of numbers which describe the number of storage blocks used in an allocation
9 area.
10
 - 11 • Blockmap — In general, the term "blockmap" refers to a map describing the status
12 of the blocks in the file system.
13
 - 14 • Snapdelete — In general, the term "snapdelete" refers to an operation that removes
15 a particular snapshot from the file system. This command can allow a storage
16 block to be freed for reallocation provided no other snapshot or the active file
17 system uses the storage block.
18
 - 19 • Snapcreate — In general, the term "snapcreate" refers to the operation of retaining
20 a consistency point and preserving it as a snapshot.
21

As described herein, the scope and spirit of the invention is not limited to any of the definitions or specific examples shown therein, but is intended to include the most general concepts embodied by these and other terms.

System Elements

Figure 1 shows a block diagram of a system for an instant snapshot.

The root block 100 includes the inode of the inode file 105 plus other information regarding the active file system 110, the active map 115, previous active file systems known as snapshots 120, 125, 130 and 135 and their respective snapmaps 140, 145, 150 and 155.

Summary The active map 115 of the active file system 110 is a bitmap associated with the in-use or free status of blocks for the active file system 110. The respective snapmaps 140, 145, 150 and 155 are active maps can be associated with particular snapshots 120, 125, 130 and 135 and an inclusive OR summary map 160 of the snapmaps 140, 145, 150 and 155. Also shown are other blocks 115 including double indirect blocks 130 and 132, indirect blocks 165, 166 and 167 and data blocks 170, 171, 172 and 173. Finally, Figure 1 shows the spacemap 180 including a collection of spacemap blocks of numbers 181, 182, 183, 184 and 190.

1 The root block 100 includes a collection of pointers that are written to the
 2 file system when the system has reached a new CP (consistency point). The pointers are
 3 aimed at a set of indirect (or triple indirect, or double indirect) inode blocks (not shown)
 4 or directly to the inode file 105 consisting of a set of blocks known as inode blocks 191,
 5 192, 193, 194 and 195.

6
 7 The number of total blocks determines the number of indirect layers of
 8 blocks in the file system. The root block 100 includes a standard quantity of data, such as
 9 128 bytes. 64 of these 128 bytes describe file size and other properties; the remaining 64
 10 bytes are a collection of pointers to the inode blocks 191, 192, 193, 194 and 195 in the
 11 inode file 105. Each pointer in the preferred embodiment is made of 4 bytes. Thus, there
 12 are approximately 16 pointer entries in the root block 100 aimed at 16 corresponding
 13 inode blocks of the inode file 105 each including 4K bytes. If there are more than 16
 14 inode blocks, indirect inode blocks are used.

15
 16 *Sub 2* In a preferred embodiment, ~~file blocks~~ are 4096 bytes and inodes are 128
 17 bytes. It follows that each block of the inode file contains 32 (i.e. $4,096/128$) separate
 18 inodes that point to other blocks 115 in the active file system.

19
 20 Inode block 193 in the inode file 105 points to a set of blocks (1, 2, 3, ..., P)
 21 called the active map 115. Each block in the active map 115 is a bitmap where each bit
 22 corresponds to a block in the entire volume. A "1" in a particular position in the bitmap

1 correlates with a particular allocated block in the active file system 110. Conversely, a
 2 "0" correlates to the particular block being unused by the active file system 110. Since
 3 each block in the active map 115 can describe up to 32K blocks or 128 MB, 8 blocks are
 4 required per GB, 8K blocks per TB.

5
 6 *Sub A37* Another inode block in the inode file 105 is inode block N 212. This block
 7 includes a set of pointers to a collection of snapshots 120, 125, 130 and 135 of the vol-
 8 ume. Each snapshot includes all the information of a root block and is equivalent to an
 9 older root block from a previous active file system. The snapshot 120 may be created at
 10 any past CP. Regardless when the snapshot is created, the snapshot is an exact copy of
 11 the active file system at that time. The newest snapshot 120 includes a collection of
 12 pointers that are aimed directly or indirectly to the same inode file 105 as the root block
 13 100 of the active file system 110.

14
 15 As the active file system 110 changes (generally from writing files, deleting
 16 files, changing attributes of files, renaming file, modifying their contents and related ac-
 17 tivities), the active file system and snapshot will diverge over time. Given the slow rate
 18 of divergence of an active file system from a snapshot, any two snapshots will share
 19 many of the same blocks. The newest snapshot 120 is associated with snapmap 140.
 20 Snapmap 140 is a bit map that is initially identical to the active map 115. The older
 21 snapshots 125, 130 and 194 have a corresponding collection of snapmaps 145, 150 and
 22 155. Like the active map 115, these snapmaps 145, 150 and 155 include a set of blocks

Sub A3 1 including bitmaps that correspond to allocated and free blocks for the particular CP when
 2 the particular snapmaps 145, 150 and 155 were created. Any active file system may have
 3 a structure that includes pointers to one or more snapshots. Snapshots are identical to the
 4 active file system when they are created. It follows that snapshots contain pointers to
 5 older snapshots. There can be a large number of previous snapshots in any active file
 6 system or snapshot. In the event that there are no snapshot, there will be no pointers in
 7 the active file system.

9 Blocks not used in the active file system 110 are not necessarily available
 10 for allocation or reallocation because the blocks may be used by snapshots. Blocks used
 11 by snapshots are freed by removing a snapshot using the snapdelete command. When a
 12 snapshot is deleted any block used only by that snapshot and not by other snapshots nor
 13 by the active file system becomes free for reuse by WAFL. If no other snapshot or active
 14 files uses the block, then the block can be freed, and then written over during the next
 15 copy-on-write operation by WAFL.

16
 17 The system can relatively efficiently determine whether a block can be re-
 18 moved using the "nearest neighbor rule". If the previous and next snapshot do not allo-
 19 cate a particular block in their respective snapmaps, then the block can be freed for reuse
 20 by WAFL. For WAFL to find free space to write new data or metadata, it could search
 21 the active map 115 and the snapmaps (140, 145, 150 and 155) of the snapshots (120, 125,

1 130 and 135) to find blocks that are totally unused. This would be very inefficient; thus it
 2 is preferable to use the active map and the summary map as described below

3

Summary 7

4 A summary map 160 is created by using an IOR (inclusive OR) operation
 5 139 on the snapmaps 140, 145, 150 and 155. Like the active map 115 and the snapmaps
 6 140, 145, 150 and 155, the summary map 160 is a file whose data blocks (1, 2, 3, ...Q)
 7 contained a bit map. Each bit in each block of the summary map. describes the alloca-
 8 tion status of one block in the system with "1" being allocated and "0" being free. The
 9 summary map 160 describes the allocated and free blocks of the entire volume from all
 10 the snapshots 120, 125, 130 and 135 combined. The use of the summary file 160 is to
 11 avoid overwriting blocks in use by snapshots.

12
 13 An IOR operation on sets of blocks (such as 1,024 blocks) of the active
 14 map 115 and the summary map 160 produces a spacemap 180. Unlike the active map
 15 115 and the summary map 160, which are a set of blocks containing bitmaps, the space-
 16 map 180 is a set of blocks including 181, 182, 183, 184, and 190 containing arrays of bi-
 17 nary numbers. The binary numbers in the array represent the addition of all the vacant
 18 blocks in a region containing a fixed number of blocks, such as 1,024 blocks. The array
 19 of binary numbers in the single spacemap block 181 represents the allocation of all
 20 blocks for all snapshots and the active file system in one range of 1,024 blocks. Each of
 21 the binary numbers 181, 182, 183, 184, and 190 in the array are a fixed length. In a pre-

5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

ferred embodiment, the binary numbers are 16 bit numbers, although only 10 bits are used.

In a preferred embodiment, the large spacemap array binary number 182 (0000001111111110=1,021 in decimal units) tells the file system that the corresponding range is relatively full. In such embodiments, the largest binary number 00001111111111 (1,023 in decimal) represents a range containing at most one empty.. The small binary number 184 (0000000000001110=13 in decimal units) instructs the file system that the related range is relatively empty. The spacemap 180 is thus a representation in a very compact form of the allocation of all the blocks in the volume broken into 1,024 block sections. Each 16 bit number in the array of the spacemap 180 corresponds to the allocations of blocks in the range containing 1,024 blocks or about 4 MB. Each spacemap block 180 has about 2,000 binary numbers in the array and they describe the allocation status for 8 GB. Unlike the summary map 120, the spacemap block 180 needs to be determined whenever a file needs to be written.

Figure 2 shows a block diagram of an instant snapshot.

The old root block 200 of snapshot #1 201 includes the inode of the inode file 202 plus other information regarding the previous active file system known as snapshot #1 201, the snapmap 205, earlier active file systems known as snapshot #2 210, snapshot #3 215 and snapshot #4 220, and their respective snapmaps 225, 230 and 235.

1

SWA5 7

The snapmap 205 of the previous active file system, snapshot #1 201, is a bitmap associated with the vacancy of blocks for snapshot #1 201. The respective snapmaps 225, 230 and 235 are earlier active maps that can be associated with particular snapshots 210, 215 and 220 and an inclusive OR summary map 245 of the snapmaps 225, 230 and 235. Also shown are other blocks 211 including double indirect blocks 240 and 332, indirect blocks 250, 251 and 252 and data blocks 260, 262, 263 and 264. Finally, Figure 2 shows the spacemap 270 of snapshot #1 201 including a collection of spacemap blocks of binary numbers 272, 273, 274, 275 and 276.

The old root block 200 includes a collection of pointers that are written to the previous active file system when the system had reached the previous CP. The pointers are aimed at a set of indirect (or triple indirect, or double indirect) inode blocks (not shown) or directly to the inode file 202 consisting of a set of blocks known as inode blocks 281, 282, 283, 284 and 285.

An inode block 281 in the inode file 202 points to other blocks 328 in the old root block 201 starting with double indirect blocks 240 and 332 (there could also be triple indirect blocks). The double indirect blocks 240 and 332 include pointers to indirect blocks 250, 251 and 252. The indirect blocks 250, 251 and 252 include pointers that are directed to data leaf blocks 260, 262, 263 and 264 of the active file system 201.

Sub A57
 1 Inode block 283 in the inode file 202 points to a set of blocks (1, 2, 3, ..., P)
 2 called the snapmap 205. Each block in the snapmap 205 is a bitmap where each bit cor-
 3 responds to a block in the entire volume. A "1" in a particular position in the bitmap cor-
 4 relates with a particular allocated block in the active file system 201. Conversely, a "0"
 5 correlates to the particular block being free for allocation in the old root block 201. Each
 6 block in the snapmap 205 can describe up to 32K blocks of 128 MB.

8 Inode file 202 also includes inode block N 285. This block includes a set of
 9 pointers to a collection of earlier snapshots, snapshot #2 210, snapshot #3 215 and snap-
 10 shot #4 220 of the volume. Each snapshot includes all the information of a root block
 11 and is equivalent to an older root block from a previous active file system.

Sub A67
 12
 13 Snapshot #1 201 also includes an old summary map 245 and old spacemap
 14 blocks 270. Although these blocks of data are included in snapshot #1 201 and previous
 15 snapshots, in a preferred embodiment, this data is not used by the active file system of
 16 figure 2.

18 *Method of Use*

20 Figure 3 shows a flow diagram of a method for using a system as shown in
 21 figure 1.

SUBA77

A method 300 is performed by the file system 100. Although the method 300 is described serially, the steps of the method 300 can be performed by separate elements in conjunction or in parallel, whether asynchronously, in a pipelined manner, or otherwise. There is no particular requirement that the method 300 be performed in the same order in which this description lists the steps, except where so indicated.

At a flow point 305, the file system 100 is ready to perform a method 300.

At a step 310, a user will request a snapshot of the file system 100.

At a step 315, a timer associated with the file system 100 initiates the creation of a new snapshot.

At a step 320, the file system 100 receives a request to make a snapshot.

At a step 325, the file system 100 creates a new file.

At a step 330, the root node of the new file points to the root node of the current active file system.

SUBA77

At a step 335, the file system 100 makes the file read only.

Sub-A87

At a step 340, the file system 100 updates the new summary map by using an inclusive OR of the most recent snapmap and the existing summary file. This step must be done before any blocks are freed in the corresponding active map block. If multiple snapshots are created such that the processing overlaps in time, the update in step 340 need only be done for the most recently created snapshot.

At a flow point 345, the snapshot create and the summary file update is completed and the snapshot creation is done.

An analogous method may be performed for snapshot delete.

Figure 4 shows a flow diagram of a method for updating a summary map.

Sub-A97

A method 400 is performed by the file system 100. Although the method 400 is described serially, the steps of the method 400 can be performed by separate elements in conjunction or in parallel, whether asynchronously, in a pipelined manner, or otherwise. There is no particular requirement that the method 400 be performed in the same order in which this description lists the steps, except where so indicated.

At a flow point 410, the file system 100 is ready to update the summary map.

SWA97

At a step 411, update of the summary map is triggered by a "snapdelete" command from an operator or user. As part of this step, the file system 100 receives and recognizes the "snapdelete" command.

At a step 412, the file system 100 responds immediately to the operator or user, and is ready to receive another operator or user command. However, while the operator or user sees a substantially immediate response, the file system 100 continues with the method 400 to process the "snapdelete" command.

At a step 413, the file system 100 marks an entry in the fsinfo block to show that the selected snapshot (designated by the "snapdelete" command) has been deleted.

At a step 414, the file system 100 examines the snapmap for the selected snapshot for blocks that were in use by the selected snapshot, but might now be eligible to be freed.

At a step 415, the file system 100 examines the snapmaps for (A) a snapshot just prior to the selected snapshot, and (B) a snapshot just after the selected snapshot. For blocks that were in use by the selected snapshot, the file system 100 sets the associated bit to indicate the block is FREE, only if both of those snapmaps show that the block was free for those snapshots as well.

1
2 The method 400 continues with the flow point 440.

3
4 *Sub A167* At a step 421, update of the summary map is triggered by a write allocation
5 operation by the file system 100. In a preferred embodiment, a write allocation operation
6 occurs for a selected section of the mass storage. The "write allocation" operation refers
7 to selection of free blocks to be seized and written to, as part of flushing data from a set
8 of memory buffers to mass storage. As part of this step, the file system 100 determines a
9 portion of the summary map corresponding to the selected section of the mass storage.

10
11 At a step 422, the file system 100 recalculates the summary map for the
12 portion of the summary map corresponding to the selected section of the mass storage.

13
14 The method 400 continues with the flow point 440.

15
16 *Sub A117* At a step 431, update of the summary map is triggered by a background op-
17 eration. In a preferred embodiment, the file system 100 updates about one 4K data block
18 of the summary map.

19
20 At a step 432, the file system 100 recalculates the summary map for the
21 portion of the summary map selected to be updated.

1 The method 400 continues with the flow point 440.

2
3 At a flow point 440, the file system 110 has updated at least a portion of the
4 summary map, and is ready to be triggered for further updates later.

5
6 Figure 5 shows a block diagram of copy-on-write maintenance of the active
7 map.

8
9 When blocks are freed in the active map, the file system 110 is careful to
10 not reuse those blocks until after a consistency point has passed (and thus that the newly
11 free status of the block has been recorded in a snapshot). Accordingly, the file system
12 110 maintains two copies of the active map, a "true" copy 501 and a "safe" copy 502.

13
14 In normal operation 510 (outside a time when a consistency point is being
15 generated), the file system 110 maintains both the "true" copy 501 and the "safe" copy
16 502 of the active map. Since in normal operation 510 blocks can only be freed, not allo-
17 cated, only changes from IN-USE to FREE are allowed. The file system 110 makes all
18 such changes in the "true" copy 501, but does not make them to the "safe" copy 502. The
19 "safe" copy 502 therefore indicates those blocks which can be safely allocated at the next
20 consistency point.

1 While generating a consistency point, during a write allocation interval 520,
2 blocks can be either freed (by continued operation of the file system 110) or allocated (by
3 the write allocation operation). Both types of change are made to both the "true" copy
4 501 and the "safe" copy 502.

5
6 While still generating a consistency point, during a flush data to disk inter-
7 val 530, blocks can again only be freed (by continued operation of the file system 110);
8 they cannot be allocated because the write allocation interval 520 is finished for that con-
9 sistency point. The file system 110 makes all such changes in the "safe" copy 502, but
10 does not make them to the "true" copy 501. At the end of the flush data to disk interval
11 530, the file system 110 switches the roles of the "true" copy 501 and the "safe" copy
12 502, so that all such changes were in fact made to the new "true" copy 501 only.

13 14 *Alternative Embodiments*

15
16 Although preferred embodiments are disclosed herein, many variations are
17 possible which remain within the concept, scope, and spirit of the invention, and these
18 variations would become clear to those skilled in the art after perusal of this application.